Final Project Report: Model of Hospital Data

Cameron Stivers, Beau McAndrew
California Polytechnic State University - Stat 334, Winter 2022

Our initial model consisted all of the possible predictor variables from the dataset: average patient age, percent chance of acquiring an infection while in the hospital, number of cultures performed per 100 patients without signs of pneumonia, whether the hospital is associated with a medical school (baseline = no), the region of the country (baseline = North Central, predictors = North East, South, West), number of beds at the hospital, average daily numbers of patients at the hospital, average number of nurses at the hospital, and the percentage of a list of medical services that are provided at the hospital to predict the response variable, the average length of stay for a patient (**Table 1A**). We immediately noticed some issues with this model.

In checking the assumptions of the model, we knew that the independence assumption was good, as the data was from a random sample of hospitals and the length of the hospital stay of one patient would not have an effect on the length of another patient's hospital stay. We noticed in the Predicted vs Residuals Plots (**Plot 1B**) that there was a violation of the linearity assumption for the Culture vs Residuals Plot as there was some curving. The linearity assumption was not violated for any of the other predictors. We did not notice any fan shaped pattern in the Predicted vs Residuals Plots (**Plot 1B**) so there were no violations of the equal variance assumption. There was a violation of the normality assumption as the points seem to slightly curve off the diagonal of the Normal Probability Plot (**Plot 1C**). We weren't sure if the curve was enough to constitute a violation of the normal assumption so we ran a Shapiro-Wilk test that yielded a p-value <.00001 to confirm that there was a normality violation. Then, we checked the multicollinearity of the predictors (**Table 1D**) and saw that there were severe multicollinearity issues in the Beds and Census predictor variables with VIFs of 35.7 and 34.2, respectively. There were not any influential points as none of the Cook's distance for the points exceeded the Cook's distance cut-off of .956 (**Plot 1E**).

To fix the violation of the linearity assumption for the Culture predictor variable, we transformed the Culture variable by taking the natural log of it (**Table 2A**). This got rid of any curving in the new Log(Culture) vs residuals plot (**Plot 2B**). To fix the normality violation, since the residuals were right-skewed, we tried two ways of transforming the y variable (LenStay) by taking the square root of the variable and taking the log of the variable. Taking the log(LenStay) did not fully fix the normality violation (**Plot 2C**) but appeared to make it better as the points did not curve off the diagonal of the QQ plot as much compared to the original QQ plot and the Shapiro-Wilk test gave us a p-value of .01402. To fix the multi-collinearity issues for the Beds and Census variables, we created two new variables, bedprop which is the proportion of beds per average number of patients at the hospital (Beds / Census) and nurseprop which is the proportion of nurses per average number of patients at the hospital (Nurses / Census). This got rid of all of the multicollinearity issues in the model (**Table 2D**).

Next, we began trying to find interactions for our model. First, we tried testing interactions between the two categorical variables of School and Region, respectively, with all of the other predictor variables, as we thought that the effect of a hospital being associated with a school on the length of the stay might change depending on the value of another predictor variable or that the effect of the region the hospital is in on the length of the stay might change depending on the value of another predictor variable. However, after going through each of these possible interactions, none of them ended up being significant. Next, we thought that age might have an effect on the length of stay based on another predictor variable, so we tried all possible interactions with age and the other predictors but again, did not find any significant results. We proceeded to go through nearly all possible interactions to find any significant interactions and we found that the interaction between InfRisk and Census was significant as the effect of the risk of infection on the average length of stay changes depending on the average daily number of patients at the hospital (**Table 3A**). We could not find any other significant interactions between predictors.

For the process of reducing the variables in the model, we used two different methods that ended up giving us the same final model. First, we used the best subsets regression method based on the lowest AIC (**Table 4A**) and were suggested with a model that had Age, Region, XRay, nurseprop, Census, InfRisk, and InfRisk * Census (**Table 4B**). This was a good sign, as the suggested model contained the interaction we had made and did not exclude either of the predictors for the interaction. We also ran a best subsets regression based on the lowest sbc and ended up with the same model suggestion. After putting the model together and running it (**Table 4B**), we were satisfied with the model we had but we were conflicted with what to do with the Age predictor variable, as it had an individual p-value just over .05. After some deliberation, we decided to remove Age from the model as it did not affect the sbc, aic, or R^2 all that much and it was insignificant. This is what we believe to be the best model (**Table 5A**).

Because the response variable for our final model (**Table 5A**) is log(LenStay), each of the predictor variables must be interpreted based on how they affect the median average length of stay for a patient. These interpretations are based on when the other predictors are adjusted for. It is not meaningful to interpret the intercept of this model, since it would not make sense for a hospital to have 0 nurses, 0 patients per day, etc., nor does our data suggest that exists. As a whole, the region of a hospital does have an effect on the median average stay length of its patients (**Table 5B**), after adjusting for the other predictors. Hospitals in the Northeast are associated with a median average stay length that is 1.083 times that of hospitals in the North Central region. Hospitals in the West are associated with a median average stay length that is 0.895 times that of hospitals in the North Central region. Hospitals in the South are associated with a median average stay length that is 0.969 times that of hospitals in the North Central region, but this difference is insignificant due to the large p-value. Each increase of 1 x-ray performed per 100 patients without signs of pneumonia multiplies the median average stay length by 1.0014. Each increase of 1 average nurse per 10 average daily patients is associated with multiplying the median average stay length by 0.992. Each increase of 1 percent in infection risk is associated with multiplying the median average stay length by 1.017. Even though this association is insignificant, it must be included since the InfRisk:Census interaction is significant. Each increase of 1 in the average number of daily patients is associated with multiplying the median average stay length by 0.999. Increasing the percent chance of acquiring an infection while at the hospital by 1 percent multiplies the effect that each increase of 1 in the average number of daily patients has on the median average stay length by 1.0002. Also, increasing the average number of daily patients by 1 multiplies the effect that increasing the percent chance of acquiring an infection by 1 has on the median average stay length by 1.002.

The adjusted R^2 of our final model is 0.6136. This means that the Region of the hospital, number of x-rays performed per 100 patients without signs of pneumonia, proportion of nurses per patients, average daily number of patients, percent chance of acquiring an infection, and the effect of the interaction between the infection risk and daily number of patients explains 61.36% of the variability in the average length of stay for a patient.

The only remaining problem in our model that we were unable to fix was the issue of the normality assumption being violated. We tried everything in our knowledge to fix the issue and were able to make it somewhat better by taking the log of the response variable (LenStay), but it still failed the Shapiro-Wilk test at a significance level of .05 with a p-value of 0.004343 (much better compared to the p-value without taking the log of LenStay: <.00001). The points did not skew off the diagonal of the QQ plot (**Plot 5C**) too much so it is not a major issue. Also, the large sample size (n = 113) means that none of the tests or confidence intervals for the slope coefficients or the regression model will be affected, just the prediction intervals may not be accurate. We hoped to get a larger adjusted R^2 than we found but after taking a while to explore new interactions and adding and removing many variables from the model, we are led to believe that this final model has the highest possible adjusted R^2 given the variables that we have to work with. No other assumptions are violated in the model as seen in the predicted vs residuals plots (**Plot 5D** - no fan shape or curving). The VIF (**Table 5E**) of Census and the interaction InfRisk*Census were high but this is not a multicollinearity issue and is somewhat expected due to the interaction. There are no influential points as none of the Cooks' distances come close to the Cook's cut-off of .933 (**Plot 5F**). Finally, we know that this model is an improvement of the original model since the P-value of the anova test comparing the two is large, and therefore the test is insignificant (**Table 5G**).